TUXEDO RNA-Seq Workflows

rnapipe.cfmaker & rnapipe

User Documentation and Tutorial

CCRIFX Bioinformatics Core

Date: October 31st 2013

Purpose

This document covers the steps involved in the analysis of Illumina RNA-Seq data using the workflows developed within the Core. There are two aims for the document. The first provides information about the workflows and how to use the workflows. The second is a demonstration to calculate expression values for genes and differential expression values between pairs of samples from an example Illumina RNA-Seq experiments

The scripts are a wrapper for the Tuxedo commands and control the workflow. There are many benefits to this (a) reproducibility (b) automation and (c) a reduction in the turnaround time for RNA-Seq workflow.

The essence of this development effort is to automate a protocol outlined in the Nature Protocols paper entitled "Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks" by Cole et al 2012 [5a].

Introduction

The CCRIFX Bioinformatics Core developed two workflows to automate RNA-Seq analysis based on the Tuxedo package [1-2]. These are: rnapipe.cfmaker.05.31.2013.pl and rnapipe.08.14.2012.pl.

The first workflow (rnapipe.cfmaker.05.31.2013.pl) generates two configuration files (RNACONF and MDCONF). These configuration files are used with the second workflow (rnapipe.08.14.2012.pl). The configuration files contain the details of sample names, sample labels and sample comparisons. The configuration files contain details about software versions and biowulf paths for executables for tophat, bowtie2, cufflinks, cuffmerge and cuffdiff and associated options, parameters and indexes. The configuration file, RNACONF, is for the tophat and cufflinks setup whilst the configuration file, MDCONF, is for cuffmerge and cuffdiff.

The second workflow (rnapipe.08.14.2012.pl) automates steps for the tuxedo pipeline (bowtie2, tophat, cufflinks, cuffmerge and cuffdiff). Unless otherwise mentioned, all parameters used are default. The user needs to have read the manuals for tophat, bowtie and cufflinks for familiarity with the versions and the options for each module of the tuxedo suite [1-3,5].

Prerequisites

a) This documentation assumes the user has read the tophat, bowtie and cufflinks manuals and the biowulf pages for tophat, bowtie and cufflinks and understands the parameters and options available for use. The documentation assumes the user knows what versions of the third-party software to use for their analysis and the desired version of the software is available on the biowulf servers [1-2]. The parameters supplied for the options, the versions, the indices and the GTFs to the two scripts (rnapipe.cfmaker.05.31.2013.pl and rnapipe.08.14.2012.pl) are the user's responsibility.

b) This documentation assumes that the user has either a Biowulf or Moab account [1-2].

c) The tutorial provided below assumes that execution is carried out on Biowulf [1-2].

d) The following packages are installed on the user's system: (tophat, bowtie, bowtie2, cufflinks, cuffmerge and cuffdiff) [1-2]. Users need to source the shell script ccrifx_sys_setup.sh to check the availability of these software and configure the software path.

Options

a) **Table 1** shows options for rnapipe.cfmaker.05.31.2013.pl

Options	Definitions
--version	Print out the workflow version information
--rnafile RNACONFOUTPUT	A *REQUIRED* FULL path to an output RNACONF file
--mdfile MDCONFOUTPUT	A *REQUIRED* FULL path to an output MDCONF file
--mdlines NUMBER_OF_LINES_IN_OUTPUT_MDCONF_FILE	This (optional) parameter specifies how many lines the output MDCONF file should have (can be 0 for no lines) and therefore in turn how many cuffmerge/cuffdiff comparisons the user is prompted for. If no value is given the default value is zero in which case no output lines will be generated
--fqdir ROOT_DIRECTORY_TO_CRAWL_FOR_FASTQ_FILES	A *REQUIRED* FULL path to a DIRECTORY that will be crawled for fastq files whose paths and sample names will be used for RNACONF file and MDCONF file generation.
--btidx BOWTIE_INDEX	An (optional) string for the bowtie reference field for the RNACONF file. Note that the index should be compatible with the specified version of bowtie
--thopts TOPHAT_OPTIONS	An (optional) string for the tophat options
--clopts CUFFLINKS_OPTIONS	A string for the cufflinks options
--btver BOWTIE_VERSION	An (optional) version of bowtie specified for use with tophat. Note that the specified version should be compatible with the bowtie index
--thver TOPHAT_VERSION	An (optional) version of tophat to use
--clver CUFFLINKS_VERSION	An (optional) version of cufflinks to use
--cmopts CUFFMERGE_OPTIONS	An (optional) string representing options for cuffmerge
--cdopts CUFFDIFF_OPTONS	An (optional) string for options for cuffdiff
--cmver CUFFMERGE_VERSION	An (optional) string for the version of cuffmerge to use
--cdver CUFFDIFF_VERSION	An (optional) string for the version of cuffdiff to use

b) **Table 2** shows options to use with rnapipe.08.14.2012.pl.

Options	Definition
--version	Print out the workflow version information
--workdir WORKDIR	The directory in which the output should appear (output directories containing output from tophat, cufflinks, cuffmerge, and cuffdiff) as well as job files (created if it doesn't already exist). It should be a full path.
--rnaconffile RNAFILE	The configuration file for running tophat and cufflinks. See information on the file format in the FILES section of this documentation. It should be a full path.
--mergediffconf MERGEDIFFCONFFILE	The pair configuration file for running cuffmerge and cuffdiff. See information on the file format in the FILES section of this documentation. It should be a full path.
--queue [Y/N]	The value of the argument should be "Y" (using queue-mode) or "N" (NOT using queue-mode), (case-insensitive). Whether the script should be run in queuemode or not? If the script is run in queue mode, then the script will create a job file calling itself with the passed parameters. Otherwise, the script is run interactively at the user's terminal.

Tutorial

In the tutorial section, the unix commands are highlighted in **bold font**. The standard output, resulting from issuing unix commands, is highlighted in red font. The dataset used in the tutorial contains eight transcriptome sequences. The reads are paired-end constructs and are located in the following directory /data/CCRIFX/TUTORIAL/QC_Jul_30_2013/rnapipe/.

1. Setting up environments

a. Use scripts from the latest release of the workflow scripts (release folders have the yyyy-mm-dd naming convention, pick the one with the most recent date). In this case, we have the latest release dated 2013-08-29 (/data/CCRIFX/RELEASE_08.29.2013/ccrifx_root_committed/bin/)
b. For general information on setting up environments, type more /data/CCRIFX/RELEASES/README.txt. Then browse the following directory /data/CCRIFX/RELEASE_08.29.2013/ccrifx_root_committed/ and read through the following documentation to familiarize yourself with content (release_notes.txt, QUICK_START.txt and WF_EXAMPLE.txt).
c. Add the environment to your path using the following command (replace the release folder with the most recent release at the time of execution)
source /data/CCRIFX/RELEASE_08.29.2013/ccrifx_root_committed/ccrifx_sys_setup.sh
d. After typing the above command, the $PATH environment variable is updated. You can check for this using the following unix commands (**echo $PATH**). The environment variable $CCRIFX_SYS_ROOT should also be defined. Check the environment variable $CCRIFX_SYS_ROOT is set by using the following unix command (**ls $CCRIFX_SYS_ROOT**). One of the benefits of setting these environments is so that you have access to the scripts plus dependencies without having to refer to them via full path (see step 2). If any error message is returned whilst setting environments – please report the issue using the issue-reporting template.

The standard output of the source command will provide a list of third party software (ie dependencies of the scripts and includes information regarding the versions of bowtie, cufflinks, cuffdiff and cuffmerge installed on biowulf). To find out what versions of software available to you on biowulf refer to reference [2]. You will need to know what version of the software from the Tuxedo pipeline that you wish the run.

2. Prepare configuration files with rnapipe.cfmaker.05.31.2013.pl

   The script rnapipe.cfmaker.05.31.2013.pl generates the RNACONF and MDCONF files to use with rnapipe.08.14.2012.pl. Having sourced the setup script as described in step 1, the rnapipe.cfmaker.05.31.2013.pl script should be on your path. To confirm that the script is on your path and the necessary environments set up, use the following unix command

```
which rnapipe.cfmaker.05.31.2013.pl
/spin1/users/CCRIFX/RELEASE_08.29.2013/ccrifx_root_committed/bin/rnapipe.cfmaker.05.31.2013.pl
```

   The unix command **perldoc rnapipe.cfmaker.05.31.2013.pl** provides the available perldoc written for the script as standard output (the output not shown). The perldoc shows fifteen options (Table 1). One of the options is **"–version**"; this option will provide the version of the script.

```
rnapipe.cfmaker.05.31.2013.pl – version

**********
WARNING : the following lists of parameters is defaulting to empty
and any resulting configuration file deserves editing after generation
 [ btidx , thopts , clopts , btver , thver , clver , cmopts , cdopts , cmver , cdver
]
************

RNAPIPE.ConfFileMaker 05.31.2013
```

The rnapipe.cfmaker.05.31.2013.pl script generates two configuration files (RNACONF.txt and MDFILE.txt). The 15 options for rnapipe.cfmaker.05.31.2013.pl relate to software parameters, software versions for tophat, bowtie, cufflinks, cuffdiff and cuffmerge. The user needs to specify the content for creation of the configuration files. Reading the manuals of the third-party software is a requirement [1-2,5].

In this tutorial, we have 8 samples comprising 4 sets of replicates. This project requires two separate comparisons. The first comparison will be replicate set 1 (sample CP {6}, sample 7MCP {3}) versus replicate set 2 (sample IRP {8}, sample 8MRP {4}) and the second comparisons will be replicate set 3 (sample CT {7}, sample 3MCT{1}) versus replicate set 4 (sample CIR {5}, sample 4MRT {2}).

There is an interactive portion of the rnapipe.cfmaker.05.31.2013.pl script where the user will need to provide the number and the type of comparisons to make. To create the configuration file for this project, we set the -mdlines option to 2 (see command below).

For this RNA-Seq analysis, bowtie version 2.0.0-beta7, tophat version 2.0.4, cufflinks version 2.0.2, cuffdiff version 2.0.2 and cuffmerge version 2.0.2 are used. In addition, the human hg19 index for bowtie2 and the UCSC GTF gene annotation file are used. Both the bowtie2 index and the GTF file are located on biowulf /fdb/igenomes/ subdirectories. The full path for the bowtie index for the organism in your RNA-Seq experiment is required. If it is not installed centrally on biowulf, you'll need to install a Bowtie index for the organism in your RNA-Seq experiment [1-2,5].

```
rnapipe.cfmaker.05.31.2013.pl \
--rnafile /data/CCRIFX/TUTORIAL/QC_Jul_30_2013/rnapipeOUT40/RNACONF.txt  \
--mdfile /data/CCRIFX/TUTORIAL/QC_Jul_30_2013/rnapipeOUT40/MDCONF.txt \
--mdlines 2 \
--fqdir /data/CCRIFX/TUTORIAL/QC_Jul_30_2013/rnapipe/ \
--btidx /fdb/igenomes/Homo_sapiens/UCSC/hg19/Sequence/Bowtie2Index/genome \
--thopts '-G /fdb/igenomes/Homo_sapiens/UCSC/hg19/Annotation/Genes/genes.gtf --
no-novel-juncs' \
--clopts '-G /fdb/igenomes/Homo_sapiens/UCSC/hg19/Annotation/Genes/genes.gtf -b
/fdb/igenomes/Homo_sapiens/UCSC/hg19/Sequence/WholeGenomeFasta/genome.fa -v'
\
--btver 2.0.0-beta7 \
--thver 2.0.4 \
--clver 2.0.2 \
--cmopts '-g /fdb/igenomes/Homo_sapiens/UCSC/hg19/Annotation/Genes/genes.gtf -s
/fdb/igenomes/Homo_sapiens/UCSC/hg19/Sequence/WholeGenomeFasta/genome.fa' \
--cdopts '--FDR 0.1' \
--cmver 2.0.2 \
--cdver 2.0.2 \
```

In the interactive mode, the script supplies the sample names sorted based in ascending order by ascii code and are assigned a number. At the first interactive prompt, supply **6,3 8,4** for the first line press return and at the second prompt supply **7,1 5,2** for the second line and press return. For explanations on desired comparisons read the earlier paragraphs in step 2.

The RNACONF file generated should have 9 tab-separated columns. The 9 fields are a sample label, a bowtie reference, filename for read 1, filename for read 2, tophat options, cufflinks options, bowtie version, tophat version, and cufflinks version. Each of the fields in the output file is filled in with user-specified parameters. Each of these parameters is optional. If no parameter is passed, the value of the parameter defaults to blank/empty.

The MDCONF file generated contains at least 8 columns and these are cuffmerge version, cuffdiff version, cuffmerge options, cuffdiff options, one or more columns with comma-separated labels for sample names of replicates and one or more columns for labels of the replicates. The number of columns (tab-separated values) is variable depending on the number of samples. The samples to be processed in a cuffmerge and cuffdiff operation are specified. The fields for cuffmerge version, cuffdiff version, cuffmerge options, and cuffdiff options are included across all lines. Each of these parameters is optional. If no value is passed, then the default value is empty/blank.

Additional operation notes:

Use single quotes to supply parameters for the following options only (thopts, clopts, cmopts, cdopts). Multiple parameters for a particular tool should be separated simply by a space, and the whole string must be enclosed with single quotes (see the options supplied in the above command for thopts, clopts and cmopts). If you do not want to specify any parameters for an option, two single quotes must be provided. For example: **--cdopts ''**. Versions of programs need not be enclosed in single quotes (btver, thver, clver, cmver, cdver). All other parameters supplied need not be enclosed in quotes. The -version parameter should be used used on it's own with the rnapipe.cfmaker script (ie no other option should be used with -version parameter).

The file names should only have four fields separated by underscores. Initially to run the fastqc workflow, the files must have 5 fields: eg HP04_post_ATCACG_L006_R1_all.fastq. This needs to be changed to have only 4 fields: eg HP04post_ATCACG_L006_R1_all.fastq

Be cognizant of how the -mdlines parameter affects configuration in terms of the number of comparisons to perform.

The parameters we use here is for no novel discovery for human data. It is an example to highlight use of the pipeline. For your own analysis – configure according to your experiment and analysis plan.

3. Running RNA-Seq analysis with the rnapipe.08.14.2012 script

   The rnapipe.08.14.2012 script can be used to run the tuxedo pipeline (tophat, cufflinks, cuffmerge and cuffdiff) on different samples with user-set versions and parameters.

   The unix environments are set up in step 1 from the same terminal session In this tutorial, we use the parameters supplied to the RNACONF and MDCONF configuration files as described in step 2. In this step, we use the rnapipe.08.14.2012 script to run the tuxedo pipeline (tophat, cufflinks, cuffmerge and cuffdiff) in parallel on different samples with user-specified versions and parameters.

```
rnapipe.08.14.2012.pl \

--workdir /data/CCRIFX/TUTORIAL/QC_Jul_30_2013/rnapipeOUT40/ \

--rnaconffile /data/CCRIFX/TUTORIAL/QC_Jul_30_2013/rnapipeOUT40/RNACONF.txt \

--mergediffconf /data/CCRIFX/TUTORIAL/QC_Jul_30_2013/rnapipeOUT40/MDCONF.txt \

 --queue Y \
```

   Check your jobs. Use the following command to see jobs submitted by the script and queued for execution using **qstat –u <username>**

   The time the job takes to run to completion will depend on several factors such as the number of sequences, the number of reads in each transcriptome, the types and the number of comparisons in the analysis and the availability of compute nodes. Expect the rnapipe analysis in this tutorial to take about 24 – 36 hours to run through all the steps.

   The RNACONF.txt file contains the configuration for tophat and cufflinks. The script (**rnapipe.08.14.2012.pl)** creates two directories (one for tophat output and one for cufflinks output) for each sample. One directory ends with "_tophatout" whilst the other ends with "_cufflinksout". The tophat and cufflinks scripts run in tandem for each sample. The two programs, as a set, run in parallel with respect to samples. TopHat aligns the reads to the genome using bowtie and determine splice sites. All the BAM files associated with these are in the "_tophatout" directory. Cufflinks assembles transcripts; the three main types of files (transcripts.gtf, isoforms.fpkm_tracking and genes.fpkm_tracking) are in the "_cufflinksout" directory. For more details about the content of these files refer to bowtie, tophat and cufflinks manuals [1,2,5a].

The MDCONF file contains the configurations for cuffmerge and cuffdiff. Cuffcompare compares transcript assemblies to annotation. Cuffmerge merges two or more transcript assemblies. Cuffdiff determines differentially expressed genes and transcripts and detects differential splicing and promoter usage. The output files (*_exp.diff, *.fpkm_tracking, *.count_tracking and *.read_group_tracking) are generated for genes, isoforms, CDS and TSS. These are located in pipe_merge _00000 directories. Examine each of these outputs in excel. For details about the content of these files refer to the cufflinks manual [1-2,5a].

A couple of relevant powerpoint presentations written by various Core members on the rnapipe workflow could be used for reference These ppt files can be found in the MeetingNotes sub directory of project 413 in svn. In the presentations – there is material on the log files, the OU files, the job files, and the directory structures that the rnapipe script generates; the presentations explain information in these files such as the commands executed and the logs detailing the progression of scripts. There is also an associated flowchart for reference.

References

1. Manuals for Bowtie2, TopHat and Cufflinks

http://bowtie-bio.sourceforge.net/manual.shtml
http://tophat.cbcb.umd.edu/manual.shtml
http://cufflinks.cbcb.umd.edu/

2. Biowulf documentation for Bowtie2, TopHat and Cufflinks

http://biowulf.nih.gov/apps/bowtie.html

http://biowulf.nih.gov/apps/tophat.html

http://biowulf.nih.gov/apps/cufflinks.html

3. Biowulf documentation for monitoring jobs on the queue

http://biowulf.nih.gov/user_guide.html
http://biowulf.nih.gov/user_guide.html#monitor

4. Unix cheat sheets

http://www.cyberciti.biz/tips/linux-unix-commands-cheat-sheets.html

5. References

a. Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley DR, Pimentel H, Salzberg SL, Rinn JL, Pachter L. (2012). Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nat Protoc. 7:562-578.

b. Trapnell C, Hendrickson DG, Sauvageau M, Goff L, Rinn JL, Pachter L. (2013). Differential analysis of gene regulation at transcript resolution with RNA-seq. Nat Biotechnol. 31:46-53.

c. Roberts A, Pimentel H, Trapnell C, Pachter L. (2011). Identification of novel transcripts in annotated genomes using RNA-Seq. Bioinformatics 27:2325-9.

d. Roberts A, Trapnell C, Donaghey J, Rinn JL, Pachter L. (2011). Improving RNA-Seq expression estimates by correcting for fragment bias. Genome Biol. 12:R22.

e. Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, Salzberg SL, Wold BJ, Pachter L. (2010). Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat Biotechnol. 28:511-5.

f. Langmead B, Trapnell C, Pop M, Salzberg SL. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. 10:R25.

g. Trapnell C, Pachter L, Salzberg SL. (2009). TopHat: discovering splice junctions with RNA-Seq. Bioinformatics 25:1105-11.

FAQ

1) What hg19 bowtie indices are available for bowtie2 on biowulf
Use bowtie2 version with Bowtie2 index
/fdb/igenomes/Homo_sapiens/UCSC/hg19/Sequence/Bowtie2Index/genome

2) What hg19 bowtie indices are available for bowtie on biowulf
Use bowtie version with Bowtie index (bowtie version 1)
/fdb/igenomes/Homo_sapiens/UCSC/hg19/Sequence/BowtieIndex/genome

3) The merged.gtf file generated by Cuffmerge (as part of rnapipe 08.14.2012) contains CUFF IDs instead of Ensembl IDs in the Ensembl ID field for hundreds of genes. These genes have a gene symbol associated with them indicating that they are a known genes, however the Ensembl Ids are absent.

As per the documentation and online forums, this is the normal manner in which novel isoforms are given intermittent IDs in the merged.gtf file. Some of these CUFF ids could be resolved at the cuffdiff stage. This is an observation that tends to occur with a known gene but novel isoform.

Contact

If you identify areas where you would like to expand the user documentation or would like to make an update to the user documentation or provide comments/feedback please contact Yvonne Edwards or Parthav Jailwala.

CCRIFX Bioinformatics Core
Advanced Biomedical Computing Center (ABCC)

Information Systems Program
Leidos Biomedical Research, Inc.
Frederick National Laboratory for Cancer Research (FNLCR)
P. O. Box B, Frederick, MD 21702
Phone: 301.594.1395
Fax: 301.480.0391
http://ccrifx.cancer.gov

Citation

If you want to cite this workflow, please use the following link:

Tuxedo RNA-Seq pipeline:
http://ccrifx.cancer.gov/apps/site/workflows_for_bioinformatics_analysis

This document was lasted updated November 29, 2013.